ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

Efficient Similarity Measure for Document Classification and Clustering Using Tf-Idf

Sony K T
Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Kothamangalam, Kerala
sonykareparambil@gmail.com

Neethu K P
Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Kothamangalam, Kerala
neethukannan911@gmail.com

Aby Abahai T
Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Kothamangalam, Kerala, India
abytom@gmail.com

***Abstract*—Measuring the similarity between documents is an important operation in the text processing field. Document clustering is an effective text mining method which classifies similar documents in to a group. Similarity Measure for Text Processing(SMTP) consider three features appears in , a) both documents, b) only one document, and c) none of the documents. The Similarity Measure with tf-idf is extended to gauge the similarity between two sets of documents. In SMTP, similar documents case is not covered that is, standard deviation for a particular feature tending to zero is not covered. In this work, proposed an efficient SMTP similarity measurement Instead of counting difference between features, our proposed system gives weightage for features using tf-idf. In this system absence and presence of a property has more important than similarity between documents features. The results show that the performance obtained by the proposed measure is better than that achieved by existing SMTP and other measures.**

***Index Terms*— Text processing, term frequency, inverse term frequency, tf-idf, SMTP, Document clustering,**

I. INTRODUCTION

TEXT processing plays an important role in information retrieval, data mining, and web search. In text processing, the bag-of-words model is commonly used. A document is usually represented as a vector in which each component indicates the value of the corresponding feature in the document. Measuring the similarity [1] between documents is an important operation in the text processing field. The feature value can be term frequency that is, the number of occurrences of a term appearing in the document, relative term frequency , the ratio between the term frequency and the total number of occurrences of all the terms in the document set, or a combination TF-IDF. The dimensionality of a document is large and the resulting vector is sparse, most of the feature values in the vector are zero. Such high dimensionality and sparsity can be a severe challenge for similarity measure which is an important operation in text processing.

Similarity Masure for Text Processing, that is SMTP[2] The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with a present feature decreases. The contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The measure is applied in several text applications, including single label classification, multi-label classification, *k*-means like clustering, and hierarchical agglomerative clustering, and the results obtained demonstrate the effectiveness of the proposed similarity measure.

The proposed system is a SMTP measure with tf-idf calculation for computing the similarity between two documents. Several characteristics are embedded in this measure. It is a symmetric measure. In SMTP, similar documents case is not covered that is, standard deviation for a particular feature tending to zero is not covered. In this work, proposed an efficient SMTP similarity measurement Instead of counting difference between features, our proposed system gives weightage for features using tf-idf.The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with a present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The measure is applied in several text applications, including single label classification, multi-label classification, *k*-means like

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

clustering, and hierarchical agglomerative clustering, and the results obtained demonstrate the effectiveness of the proposed similarity measure.

II. RELATED STUDIES

A lot of measures have been proposed for computing the similarity between two vectors. Euclidean distance [3] is a well-known similarity metric taken from the Euclidean geometry field. Manhattan distance [4], similar to Euclidean distance and also known as the taxicab metric, is another similarity metric. The Canberra distance metric is used in situations where elements in a vector are always non-negative.

[5] proposed a system which uses intermediate features of maximum overlap wavelet transform (IMOWT) as a pre-processing step. The coefficients derived from IMOWT are subjected to 2D histogram Grouping. This method is simple, fast and unsupervised. 2D histograms are used to obtain Grouping of color image. This Grouping output gives three segmentation maps which are fused together to get the final segmented output. This method produces good segmentation results when compared to the direct application of 2D Histogram Grouping. IMOWT is the efficient transform in which a set of wavelet features of the same size of various levels of resolutions and different local window sizes for different levels are used. IMOWT is efficient because of its time effectiveness, flexibility and translation invariance which are useful for good segmentation results. The Jaccard coefficient [6] is a statistic used for comparing the similarity of two sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The Hamming distance[7] between two vectors is the number of positions at which the corresponding symbols are different. The extended Jaccard coefficient and the Dice coefficient retain the sparsity property of the cosine similarity measure while allowing discrimination of collinear vectors.

Similarity measures have been extensively used in text classification and clustering algorithms. A cosine-based pairwise adaptive similarity [8] for document clustering used cosine to calculate a correlation similarity between two projected documents in a low-dimensional semantic space and performed document clustering in the correlation similarity measure space. There is a divisive information-theoretic feature clustering algorithm for text classification using the Kullback-Leibler divergence. Euclidean distance is usually the default choice of similarity-based methods, *k*-NN [9] and *k*-mean [10].

III. DISCUSSION OF SMTP

Properties of Similarity for Text Processing are given:

1) The presence or absence of a feature is more essential than the difference between the two values associated with a present feature

2) The similarity degree should increase when the difference between two non-zero values of a specific feature decreases.

3) The similarity degree should decrease when the number of presence-absence features increases.

Similarity between Two Documents

For two documents document1 =< document11, document12, ... , document1m > and document2 = < document21, document22, . . . , document2m >. Define F as follows:

$$F(\mathbf{d}_1, \mathbf{d}_2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N_\cup(d_{1j}, d_{2j})} \quad (1)$$

Similarity measure in SMTP [2], for document1 and document2

$$S_{SMTP}(\mathbf{d}_1, \mathbf{d}_2) = \frac{F(\mathbf{d}_1, \mathbf{d}_2) + \lambda}{1 + \lambda}. \quad (2)$$

Where,

$$N_*(d_{ik}, d_{jk}) = \begin{cases} 0.5 \times \left(1 + \exp\left\{-\left(\frac{d_{ik} - d_{jk}}{\sigma_k}\right)^2\right\}\right) \\ \quad if\ d_{ik}d_{jk} > 0 \\ 0, \quad if\ d_{ik} = 0\ and\ d_{jk} = 0 \\ -\lambda,\ otherwise. \end{cases} \quad (3)$$

$$N_\cup(d_{1j}, d_{2j}) = \begin{cases} 0, & \text{if } d_{1j} = 0 \text{ and } d_{2j} = 0 \\ 1, & \text{otherwise.} \end{cases} \quad (4)$$

IV. PROPOSED METHODOLOGY

In this section the SMTP measure is used, for clustering Similarity Measure for Text Processing with TF-IDF calculation is introduced.

SMTP measure takes into account the following three cases:
a) The feature considered appears in both documents,
b) The feature considered appears in only one document, and
c) The feature considered appears in none of the documents.

In SMTP, similar documents case is not covered that is, standard deviation for a particular feature tending to zero is not covered Let us consider there are two similar documents d1=<1,1,1> and d2 =<1,1,1>with three similar features.

$$F(d_1, d_2) = \frac{0.5 \times \left(1 + \exp\left\{-\left(\frac{1-1}{0}\right)^2\right\}\right) + 0.5 \times \left(1 + \exp\left\{-\left(\frac{1-1}{0}\right)^2\right\}\right) + 0.5 \times \left(1 + \exp\left\{-\left(\frac{1-1}{0}\right)^2\right\}\right)}{1+1+1} \quad (5)$$

Since $\exp^{\{-((1-1)/0)^2\}}$ is not a number (or not defined) considering this part as 0 will give

$$F(d_1, d_2) = \frac{0.5 \times (1) + 0.5 \times (1) + 0.5 \times (1)}{1+1+1} = \frac{1.5}{3} = 0.5 \quad (6)$$

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

$$S_{SMTP}(d_1, d_2) = \frac{0.5+1}{1+1} = 0.75. \quad (7)$$

Similarity technique like Cosine for the similar pair of documents is given

$$S_{Cos}(d_1, d_2) = \frac{1\times1+1\times1+1\times1}{\sqrt{1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2}} = 1. \quad (8)$$

So it seems that the condition where the feature is exactly the same (or in other words standard deviation is zero for a feature) is not covered in the SMTP similarity technique. In order to overcome this limitation one supplementary condition is suggested for the SMTP similarity technique as given

$$N_*(d_{ik}, d_{jk}) = \begin{cases} 1, \text{ if } d_{ik} = d_{jk} \text{ and } d_{ik}, d_{jk} > 0 \\ 0.5 \times \left(1 + \exp\left\{-\left(\frac{d_{ik}-d_{jk}}{\sigma_k}\right)^2\right\}\right) \text{if } d_{ik}d_{jk} > 0 \\ 0, \text{ if } d_{ik} = 0 \text{ and } d_{jk} = 0 \\ -\lambda, \text{ otherwise.} \end{cases} \quad (9)$$

Now the SMTP similarity between the pair of documents is

$$F(d_1, d_2) = \frac{1+1+1}{1+1+1} = \frac{3}{3} = 1.0 \quad (10)$$

$$S_{SMTP}(d_1, d_2) = \frac{1.0+1}{1+1} = 1.0. \quad (11)$$

A. TF-IDF

A document is usually represented as a vector in which each component indicates the value of the corresponding feature in the document. The feature value can be

a. term frequency (the number of occurrences of a term appearing in the document)
b. relative term frequency (the ratio between the term frequency and the total number of occurrences of all the terms in the document set) or
c. tf-idf (a combination of term frequency and inverse document frequency)

TF= (number of time a feature in a document/total number of feature in that document)
IDF=log (total number of document/number of document contain the feature)

Algorithm 1:Tf-Tf

```
1: class Mapper
2:     method Map((docId, N),(term, o))
3:         for each element ∈ (term, o)
4:             write(term, (docId, o, N))
5:
6:  class Reducer
7:      method Reduce(term, (docId, o, N))
8:          n=0
9:          for each element     (docId, o, N) do
10:             n = n+1
11:         tf = o/N
12:       idf = log(|D| / (1+n))
13:         return (docId, (term, tf * idf))
```

TF-IDF equation:

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i
tfik = frequency of term T_k in document D_i
idfk = inverse document frequency of term T_k in C
N = total number of documents in the collection C
nk = the number of documents in C that contain T_k
idfk = log(N/nk)$\frac{N}{n_k}$

B. Clustering And Classification

This system use KNN classification and K-means Algorithm for text classification and clustering

Classification

Classification models predict categorical class labels

Two step process

i) Learning step
ii) Classification step

Learning step: - where a classification model is constructed

Classification step: - where the model is used to predict class labels for given data

KNN classification

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors

If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

The neighbors are taken from a set of objects for which the class is known

K-Means Algorithm

k-means is one of the most popular methods which produce a single clustering. It requires the number of clusters, k, to be specified in advance. Initially, k clusters are specified.

step1: each document in the document set is re-assigned based on the similarity between the document and the k clusters. Then the

step2: k clusters are updated. Then

step3: all the documents in the document set are re-assigned.

step4: process 2,3 is iterated until the k clusters stay unchanged.

V. EXPERIMENTAL RESULTS

The dataset used in this system is Reuter-8. It is the most widely used test collection for text categorization research, though likely to be superceded over the next few years by RCV1. RCV1. The RCV1 data set consists of 804414 news stories produced by Reuters from 20 Aug 1996 to 19 Aug 1997. There are 47236 features and 101 categories involved in this data set.. The data set we use contains 30000 documents, of which

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

15000 were pre designated for training and the rest were pre designated for testing. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the construe text categorization system. we investigate the effectiveness of our proposed similarity measure SMTP. The investigation is done by applying our measure in several text applications, including *k*-NN based single-label *k*-means clustering (*k*-means) [9]. We also compare the performance of SMTP with that of other five measures, Euclidean , Cosine , Extended Jaccard (EJ) , Pairwise-adaptive (Pairwise) , and IT-Sim.

k-means is one of the most popular methods which produce a single clustering. It requires the number of clusters, *k*, to be specified in advance. Initially, *k* clusters are specified. Then each document in the document set is re-assigned based on the similarity between the document and the *k* clusters. Then the *k* clusters are updated. Then all the documents in the document set are re-assigned. This process is iterated until the *k* clusters stay unchanged.

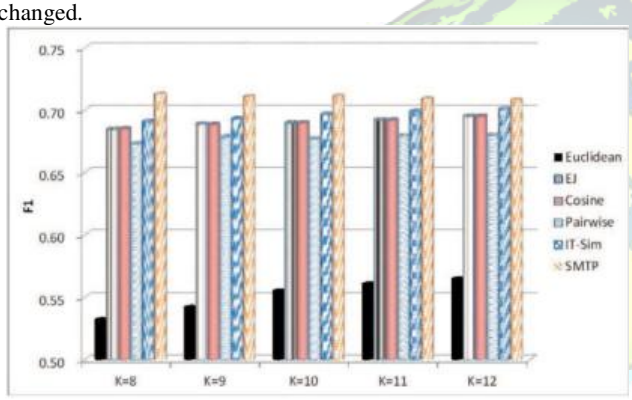

Figure 1: performance of SMTP measure with other similarity measure

Table 1: efficiency of K-means on testing data in tf-idf value with different measure

WebKB ($k = 16$)					
Euclidean	EJ	Cosine	Pairwise	IT-Sim	SMTP
75	128	106	1103	684	1084
Reuters-8 ($k = 16$)					
Euclidean	EJ	Cosine	Pairwise	IT-Sim	SMTP
258	327	305	4036	2320	1631

Table 2: classification accuracy of SL-KNN with different measure on testing data of reuter_8 in Word-Count

	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$	$k=11$	$k=13$	$k=15$
Euclidean	0.8799	0.8908	0.8872	0.8785	0.8821	0.8831	0.8799	0.8762
EJ	0.9137	0.9319	0.9315	0.9274	0.9246	0.9287	0.9296	0.9296
Cosine	0.9137	0.9265	0.9319	0.9287	0.9274	0.9278	0.9246	0.9296
Pairwise	0.9013	0.9191	0.9242	0.9223	0.9233	0.9214	0.9159	0.9159
IT-Sim	0.8963	0.9159	0.9333	0.9434	0.9411	0.9392	0.9402	0.9406
SMTP	**0.9338**	**0.9411**	**0.9420**	**0.9447**	**0.9461**	**0.9466**	**0.9434**	**0.9443**

Experimental Results show the efficiency and classification accuracy of SMTP is better than the other similarity measure algorithms. And SMTP with TF-IDF has more efficiency compared to word count

VI. CONCLUSION

There are various similarity measure and clustering techniques available with varying attributes which is suitable for find similarity between features with SMTP measurement. it provide more important for absence and presence of feature than difference between feature. In this paper, Document Similarity Measure for Classification and Clustering Using Tf-Idf is chosen because the Term Frequency- Inverse term frequency for feature weightage is calculated. Measuring the similarity between the document pairs in not considered in SMTP. In this work more efficient SMTP similarity measurement is proposed. The results have shown that the performance obtained by the proposed measure is better than that achieved by other measures.

REFERENCES

[1] [Online]. Available: http://web.ist.utl.pt/~acardoso/datasets/

[2] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering". IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014

[3] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1217–1229, Sept. 2010.

[4] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2008

[5] Christo Ananth, A.S.Senthilkani, S.Kamala Gomathy, J.Arockia Renilda, G.Blesslin Jebitha, Sankari @Saranya.S., "Color Image Segmentation using IMOWT with 2D Histogram Grouping", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 3, Issue. 5, May 2014, pp-1 – 7

[6] V. Lertnattee and T. Theeramunkong, "Multidimensional text classification for drug information," *IEEE Trans. Inform. Technol. Biomed.*, vol. 8, no. 3 pp. 306–312, Sept. 2004.

[7] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006.

[8] D. D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, Apr. 2004.

[9] J. Kogan, M. Teboulle, and C. K. Nicholas, "Data driven similarity measures for *k*-means like clustering algorithms," *Inform. Retrieval*, vol. 8, no. 2, pp. 331–349, 2005.

[10] T. Kanungo *et al.*, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2006.